Overview of Studies on Automated Chinese Essay Scoring in Recent 20 Years

Tianhuan Ma

College of Chinese Language and Culture, Jinan University, Guangzhou, China

Keywords: Chinese, Automated scoring, Essay

Abstract: This paper analyzes the basic research and application of Chinese essay automatic scoring in recent 20 years (2000-2020) in China, finding that on the whole, the Chinese essay automatic scoring system basically reaches the same level as human scorers, and the existing scoring system has been put into use in large-scale essay examination scoring in some areas. As is reflected by basic research and the working principle of machine scoring, the automatic evaluation on the language quality of essays has achieved good results, but the quality of machine-based evaluation on the semantic content and textual structure is far from satisfactory, and the underlying motivations are not that justifiable. One reason is that the basic studies on essay scoring are small in number, and another reason is that the machine-driven analysis of Chinese textual semantics remains defective. Therefore, it is crucial to realize the automation evaluation on the semantic content and textual structure of Chinese essays.

1. Introduction

Manual essay scoring has always been facing difficulty in language teaching and testing because it is time-consuming and inefficient. In the last century, studies centered around automatic essay scoring rose in response to this tricky problem. After decades of exploration, some automatic scoring systems have been applied to large-scale language testing at home and abroad, including GRE, TOEFL, Chinese essay in college entrance examination and so on. As is reflected by the development in recent years, the automation of scoring has become one of the general trends of educational measurement in the future. [1] Chinese scholars Zhou Jianshe et al.[2] also demonstrate the effectiveness of language intelligent evaluation technology and emphasize the importance of developing language intelligent evaluation. In addition, many other researchers such as Zhu Bo et al. [3] also recognize the broad prospect and huge market demand of this subject.

Therefore, this paper makes an overview of the studies on automatic scoring of Chinese essay by Chinese scholars in recent 20 years (2000-2020). It firstly sorts out the basic studies on automatic scoring, then investigates the application of existing scoring systems in the educational field, and finally proposes the research outlook of this issue in China based on the existing studies on automatic scoring in the west.

2. Basic Studies on Chinese Essay Automatic Scoring

By referring to the research ideas of automatic scoring for essays of foreign languages, scholars working on Chinese studies have produced quite a few basic studies on machine scoring, especially on feature engineering digging into the essay quality, as is shown in the following:

(1) Language: the total number of words, vocabulary level, the number of grammatical errors, the number of words beyond the syllabus, the number of associated words [4-6].

(2) Structure: positions of paragraphs, sentence sequence number, guide words[7]; connection, thematic structure, thematic progression pattern [8].

(3) Content: the number of affective words, the number of changes in affection, ratio of positive and negative affective words.

It can be seen that researchers have tried to extract a variety of features that could predict the essay quality from different dimensions. Generally speaking, studies in the early days focused on the shallow features of words, phrases and sentences, and in recent years, researchers has begun to

  DOI: 10.25236/icembe.2020.056

explore the development trend of deep and multi-index evaluation. In addition to the studies on different feature types shown in the table above, some other researchers work on the application and verification of technical methods, such as Ma Hongchao [9], showing a trend of development in depth.

An obvious defect of existing studies is that most researchers try to display the quality of essays through quantitative features, like characterizing and evaluating the theme of essays by means of subject words, which is far from enough. At the same time, compared with studies on the language quality of essays, studies on the automatic evaluation of essay content are small in number.

3. Application of Chinese Essay Automatic Scoring

With regard to the automatic scoring system for Chinese essays written by native students, the intelligent product designed by iFLYTEK has been put into use in high school and college entrance examinations in some areas of China, with the evaluation dimensions including textual structure, theme, language quality, vocabulary and plagiarism detection, which shows great reliability, objectivity and economical efficiency. He Yisong et al. [10] select the calibration sets by adopting the method of “expert random extraction & intelligent selection & cluster subsection supplement” and find that under this scoring model, the correlation coefficient for machine score and expert score is 0.939, with an overall consistency of 97.13%. It can be argued that the application of automatic scoring for Chinese essays are quite optimistic at present, and it is gradually put into large-scale use in examinations.

As for the automatic scoring system for Chinese essays written by Chinese language learners, Xu Changhuo et al. [11] tried to design a semi-automatic scoring system to recognize incorrect sentences, and pointed out that the tool was especially designed for the essays written by Chinese language learners with the elementary level of proficiency in Chinese. Obviously, this system has quite many limitations. Chinese Grammatical Error Diagnosis (CGED), introduced by Rao Gaoqi [12], aimed at automatically discovering and diagnosing syntactic errors in Chinese second language learners' essays. In addition, the scoring systems designed for essays written in Chinese as a second language developed by researchers in Singapore and Taiwan mainly focuses on the detection of grammatical errors with no consideration of semantic analysis. On the whole, little attention has been paid to the quality of essay content and the layout of essays.

On the other hand, some unreasonable problems exist in terms of the process of essay evaluation and the motivation of machine scoring. At present, the neural network method is widely used, and it works by treating the input essays as data series, and then the machine only responds to the stimulus set in the program. In other words, the quality of essays is judged through the shallow quantitative features. For example, the content of the essays is evaluated by means of the number of keywords or target words [13]. As is seen in the previous studies, the richness of vocabulary is designed as the measurement of the substantiality of content, while the “emotional sincerity” is reflected by specific vocabulary features[14]. Besides, the fluency of essays is characterized by the number of words [15].

Therefore, the most striking disadvantage of machine scoring system at present lies in its inability to analyze textual semantics. An essay is no more than a bunch of disordered words for the computer, and the quality of the essay can only be evaluated by some quantitative features. In fact, these features have little predictability for manual scoring [16]. It can be argued that the method of machine scoring and its judgment basis are completely different from the process of our appreciating an article, and the scoring process can be hardly explained.

In a word, some achievements have been made in the preliminary exploration and the construction of application platform as to the automatic scoring of Chinese essays, but the studies were barely in embryo. The existing system mainly focuses on the evaluation of language, but little effort has been made as to the studies on the objective evaluation of semantic content and textual structure.

4. Achievements and Shortcomings of Previous Studies

4.1 Achievements

(1) Basic theoretical studies on essay scoring

Article science, rhetorical studies and textual linguistics have provided important basic for the evaluation of text quality and the formulation of essay scoring standards. Among them, mature achievements in the fields of vocabulary, sentence meaning and textual semantics also offer valuable insights.

(2) Applied studies on automatic essay scoring

Some relatively mature scoring systems have been put into use in China, and great achievements have been made. At the same time, researchers are gradually shifting their focus on the linguistic level to the semantic level, attempting to further improve the accuracy and validity of scoring. Studies on essay scoring of Chinese as a second language in recent years mainly focus on the automatic assessment of language quality, such as the accuracy of vocabulary and sentences.

4.2 Shortcomings

Firstly, basic studies on the content evaluation of essays written in Chinese as a second language are small in number.

The existing basic studies on manual scoring and automatic scoring of essays primarily focus on the language level of essays, devoted to evaluate the quality of essays through such indicators as the accuracy and complexity of vocabulary and sentences. These studies have yielded fruitful results, with relatively few considerations of the quality of essay content.

At present, western researchers [17] have begun to design a more detailed rating scale for the evaluation of content of second language essay writing, but little effort has been made as to essays written in Chinese as a second language. It can be seen from the large number of existing literature on Chinese learners' essays that not many researchers pay attention to the content of essays and the evaluation methods, and even fewer researchers have been working on automatic scoring of essay content.

Secondly, basic studies on Chinese textual semantics serving for automatic scoring are not sufficient. Theoretical studies on textual semantics are scarce at present, which constitutes one of the bottlenecks restricting natural language processing [18]. Great attention should be paid to the potential and value of research on textual semantics, which is yet to be further explored.

Besides, from the perspective of practical needs of automatic essay scoring, one of the problems to be overcome at present is that machines have to understand the semantic meaning of articles. To achieve this goal, sufficient studies on textual semantics are needed to provide support for the development of scoring system. The underlying reason is that the machine must have a good mastery of semantic knowledge when making semantic analysis on articles. However, most previous studies on semantics without perspectives on computer technology can’t be directly applied to machine-based development of scoring systems, which results in the absence of application-oriented studies. In addition, the program designed for essay scoring needs to be supported by a large sea of language facts, so as to dig out the resources that can meet practical needs. However, studies in this field in terms of Chinese linguistics still fall short.

At present, one of the defects of machine scoring is that the it is confined to the shallow evaluation of the language features of essays, but fails to accurately grasp the content of essays and to reasonably assess the quality of essays. In other words, machine is still unable to judge the content, theme, logic and other issues that depend on the in-depth semantic analysis. This is what restricts the studies and application development of automatic scoring. The neural network method widely used at present works by treating essays as data series without considering the content of essays, and it analyzes essays with no justifiable basis. This is another defect that needs to be improved.

In a word, automatic essay scoring, as a major issue in the field of modern education evaluation, is the only route to the scientization of evaluation. It is one of the subsidiary subjects of artificial intelligence. Automatic scoring as a subsidiary subject of AI requires researchers to generalize rules

from language facts, which is a key to scoring. This also confirms what the author argues about the research status, that is, studies on automatic scoring of Chinese essays have just started, and many issues remain to be solved in the future.

References

[1] Randy, E.B. Educational Assessment: What to Watch in a Rapidly Changing World. Educational Measurement and Evaluation, no. 03, pp. 3-14, 2019.

[2] Zhou, J.S., Zhang, K., Luo, Y., et. al. Theoretic Study of Language Intelligence Evaluation and Its Technology Applications: Taking the English Writing Intelligent Evaluation System as an Example. Language strategy research, no. 05, pp. 12-19, 2017.

[3] Zhu, B., Fu, R.J., Sheng, Z.C., et. al. The exploration of artificial intelligence and its application in educational examination assessment evaluation. Electronic Test, no. 14, pp. 5-9, 2019.

[4] Ren, C.Y. Exploratory Research on Objective Scoring of HSK Composition. Chinese Language Learning, no. 06, pp. 58-67, 2004.

[5] Li, Y.N. Automated Assay Scoring for Testing Chinese as a Second Language. Beijing Language and Culture University, 2006.

[6] Huang, Z., Xie, J.L., Xun, E.D. feature selection for Automated Essay Scoring in HSK. Computer Engineering and Applications, no. 06, pp. 118-122, 2014.

[7] Chen, L.Y. Research on Key Techniques of Automated Chinese Essay Scoring Based on Regression Analysis. Harbin Institute of Technology, 2016.

[8] Chen, S. Correlations Between Cohesions, Theme Structures and HSK Composition Scores. Jiangsu Normal University, 2018.

[9] Ma, H.C., Guo, L., Peng, H.L. Comparison of Automatic Scoring Effect of Writing Based on SVM and BP Neural Network. Examinations Research, no. 05, pp. 8-13, 2018.

[10] He, Y.S., Sun, Y.Y., Zhang, K., et. al. Research on Selection and Optimization of Calibration Set in Computer Intelligent Scoring. China Examinations, no. 01, pp. 30-36, 2020.

[11] Xu, Y.L., Liu, H.T., Liu, Z.G. On the Theories and Methods of Language Research. Foreign Language Teaching and Research, no. 01, pp. 3-11, 2020.

[12] Rao, G.Q. Foreigner' Chinese is Good or not, AI Has the Final Say. Science Daily, no. 03, pp. 02, 2007.

[13] Ericsson, P.F. The meaning of meaning: Is a paragraph more than an equation. Logan: All USU Press Publications, 2006, pp. 28-37.

[14] He, Y.S., Sun, Y.Y., Wang, Z.L., et. al. The Exploration and Practice of Automatic Essay Scoring Based on AI for Scoring Checking in Large-scale Examinations. China Examinations no. 06, pp. 63-71, 2018.

[15] Wang, X.Z. Research on the Technical Progress of Automated Essay Scoring in Educational Assessment. Educational Measurement and Evaluation, no. 05, pp. 31-37, 2018.

[16] Bai, L.F., Wang, J. Underlying Causes for Human-machine Score Differences. Foreign Language Testing and Teaching, no. 3, pp. 44-54, 2018.

[17] Kuiken, F., Ineke, V. Functional adequacy in L2 writing: Towards a new rating scale. Language Testing, vol. 34, no. 3, pp. 321-336, 2017.

[18] Ye, F. Discourse Semantics, Shanghai: World Book Publishing Company, 2017, pp.11-23.